MCMC-free adaptive Bayesian procedures using random series prior

Weining Shen and Subhashis Ghosal *
Department of Statistics, North Carolina State University

Abstract

We consider priors for several nonparametric Bayesian models which use finite random series with a random number of terms. The prior is constructed through distributions on the number of basis functions and the associated coefficients. We derive a general result on the construction of an appropriate sieve and obtain adaptive posterior contraction rates for all smoothness levels of the function in the true model. We apply this general result on several statistical problems such as signal processing, density estimation, nonparametric additive regression, classification, spectral density estimation, functional regression etc. The prior can be viewed as an alternative to commonly used Gaussian process prior, but can be analyzed by relatively simpler techniques and in many cases allows a simpler approach to computation without using Markov chain Monte-Carlo (MCMC) methods. A simulation study was conducted to show that the performance of the random series prior is comparable to that of a Gaussian process prior.

1 Introduction

Bayesian methods have been widely used in the nonparametric statistical literature. In recent years, there is a lot of development on asymptotic properties of posterior distributions. General results have been established by Ghosal et al. [1999] for posterior consistency and by Ghosal et al. [2000] and Shen and Wasserman [2001] for posterior convergence rates. It is well known

*Research is partially supported by NSF grant number DMS-1106570.

that the optimal convergence rate for the estimation of functions is determined by the smoothness level of functions, e.g., the optimal rate of estimating a univariate α-smooth function is $n^{-\alpha/(2\alpha+1)}$ [Stone, 1982], where n is the sample size. Since the smoothness parameter α is usually unknown in practice, a natural Bayesian procedure will consider assigning a prior distribution on α. It is then of interest to investigate if the resulting mixture prior leads to posterior distributions that has optimal posterior convergence rates simultaneously for all values of α. Such procedures are called rate-adaptive.

Progress has been made in studying rate-adaptive procedures. Belitser and Ghosal [2003] considered the problem of estimating a signal with Gaussian white noise and showed that the posterior rate automatically adapts to the unknown smoothness condition if the smoothness parameter only takes values in a discrete set. Ghosal et al. [2003, 2008] showed that appropriate mixture of certain priors, such as those based on spline expansions, yield optimal posterior rates for a countable range of smoothness parameters for density estimation. A similar work has been done in a nonparametric regression setting using wavelets in Huang [2004]. Scricciolo [2006] obtained adaptive rates for density estimation problems when the smoothness parameter belongs to a discrete set. The basic idea behind this approach is to use the optimal dimension $J_{n,\alpha}$ of the model for a given smoothness level α and sample size n obtained from some appropriate rate equation to construct "optimal priors" $\Pi_{n,\alpha}$ for each α, and then mix countably many of them to construct the mixture prior which adapts for all these countably many smoothness levels. Alternatively, van der Vaart and van Zanten [2009] constructed a prior based on a randomly rescaled smooth Gaussian process, which automatically adapts for a continuous range of smoothness parameters.

In nonparametric Bayesian literature, a prior on a function is usually constructed via a stochastic process, (e.g., a Gaussian process prior in van der Vaart and van Zanten [2008, 2009]) or by expanding a function in a basis of functions or by convoluting a kernel with a random measure. In our study, we focus on the basis expansion approach by putting a prior on the coefficients of basis functions and the dimension of the basis function spaces. There are several advantages of using a prior directly on model dimension J rather than on the smoothness level. First, this bypasses the need for specifying the optimal dimension $J_{n,\alpha}$ in practice, which are given by posterior convergence theorem only up to a constant multiple. In a sense, this is more natural from a Bayesian point of view and is a common method used by practitioners. More importantly, assigning a prior directly on J allows us to obtain adaptation for all smoothness levels

in an interval rather than for only a countable number of smoothness levels.

A similar idea was used in Babenko and Belitser [2010] for the white noise models, which is equivalent to an infinite dimensional normal model. They introduced the idea of the oracle dimension for every parameter value defined by the minimizer of risk in a class of estimation problems induced by the dimension. The oracle dimension is used to slice the parameter space into different smoothness levels. They assigned a prior distribution on the oracle dimension and the projection of the infinite dimensional mean vector on the finite dimensional oracle. They showed that the risk of the Bayes estimator satisfies some desirable oracle inequalities, which lead to a complete adaptation of a Bayes estimator. Interestingly, the oracle inequality gives adaptation simultaneously for many different smoothness classes. While finishing writing this article, we also became aware of two very recent works which also use this same idea of directly assigning a prior distribution on the number of terms in a basis expansion. Rivoirard and Rousseau [2012] exclusively considered the density estimation problem using wavelet basis. de Jonge and van Zanten [2012] considered a general class of inference problems using spline basis and Gaussian priors on coefficients, and hence the resulting priors are mixtures of finite dimensional Gaussian processes. We formulate one general theorem in an abstract setting suitable as a prelude for many different inference problem where we allow arbitrary basis functions and arbitrary multivariate distributions on the coefficients of the expansion. Thus the resulting process induced on the function need not be Gaussian, and can accommodate from bounded support to heavy tailed distributions. The resulting rate obtained in the abstract theorem depends on the smoothness of the underlying function, approximation ability of the basis expansion used, tail of the prior distribution on the coefficients, prior on the number of terms in the series expansion, prior concentration and the metrics being used. The general theorem then gives rise to adaptive posterior convergence rates for different estimation problems.

We illustrate the implication of the abstract theorem for the white noise model, density estimation on compact interval or the real line, nonparametric normal, Poisson and binary regression, spectral density estimation of a stationary time series and functional regression model using the B-spline basis as one main example. B-splines have been well studied by mathematicians [de Boor, 2001] and have been used in statistics as well. Non-negativity, near orthogonality, summing to one and arbitrarily high smoothness level of B-spline functions are collectively the reasons for the popularity of B-spline basis. The idea is to approximate a function that is of interest as a linear combination of the spline

functions. Then the estimation of the function becomes equivalent to the estimation of the coefficients in the B-spline basis expansion [Truong et al., 2005]. In some cases, we do not approximate the true function directly. Instead, we consider a transformation that is needed to satisfy certain constraints. For example, for density estimation, an exponential transformation along with a normalization step is often used as the functions are required to be nonnegative and integrate to one [Stone, 1990, Ghosal et al., 2000]. We also discuss adaptation on Soblev and Besov spaces besides commonly used Hölder spaces to quantify smoothness. For nonparametric normal regression, we also remove a commonly used condition that the variance is bounded away form zero. Further, for the density estimation problem, we allow a flexible general link function instead of only the exponential link. By restricting coefficients to $(0, \infty)$, we do not even need to use a link function. The additional flexibility may be valuable in finding computational algorithms.

The random series expansion prior may be regarded as a valuable alternative to the Gaussian process prior. The Gaussian process has received a lot of attention in the literature on all aspects, from modeling a prior distribution [Leonard, 1978, Lenk, 1988] to computation [Tokdar, 2007] to asymptotics [Tokdar and Ghosh [2007], Ghosal and Roy [2006], Choi and Schervish [2007], van der Vaart and van Zanten [2007, 2008, 2009], Castillo [2008, 2012]], to applications in spatial statistics [Banerjee et al., 2008] and elsewhere. Asymptotic properties of posterior distributions based on Gaussian process priors are primarily driven by the structure of its reproducing kernel Hilbert space. While the elegant result of van der Vaart and van Zanten [2009] established that appropriately randomly rescaled Gaussian processes lead to posterior that automatically adapts to the unknown smoothness, in this paper we show that the same property also holds for random series priors using relatively elementary techniques. Computationally, Gaussian processes are relatively difficult to deal with. Except for Gaussian Markov random fields for which the integrated nested Laplace approximation method has been developed [Rue et al., 2009], the general approach to computation is to approximate the given Gaussian process by one that is generated by finitely many normal variable, obtained by conditioning the original process at a number of knots, which needs to be sufficiently large [Tokdar, 2007]. The approach adaptively chooses the knots through a reversible jump Markov chain Monte-Carolo (RJMCMC). While RJMCMC can also be used for random series prior, we came up with a method that poses a conjugate-like prior for the model and hence avoids the use of MCMC as the posterior can be represented analytically. When the sample size

n is relatively small (e.g. $n = 10$), the exact values can be obtained. When the sample size is large, we use a direct sampling strategy. Thus at least conceptually, the random series prior gives rise to a more straightforward approach to computation. It may be noted that Gaussian process and random series priors are intimately related in two ways – normal prior on the coefficients of a random series give Gaussian processes, and Karhunen-Loève expansion of a Gaussian process express it as a random series with basis consisting of eigenfunctions of the covariance kernel of the Gaussian process. Thus random series may be regarded as a more general and flexible alternative to Gaussian process that allows a more straightforward approach to computation and asymptotics.

There are two key steps in proving Bayesian adaptation results. The first is to construct a sieve that contains the true underlying model while its size is well controlled. The size is usually characterized by the existence of tests or illustrated by entropy calculation [Ghosal et al., 2000]. The other step is to construct an approximation of the true function while its approximation accuracy increases appropriately with the increasing level of smoothness. We derive a general result on the existence of an appropriate sieve that can later be used in showing asymptotic results. In particular, we illustrate its use on a few statistical models using B-spline functions as the approximation technique.

The paper is organized as follows. We introduce some notations in Section 2. In Section 3, we present main theorems of random series and B-spline functions. In Sections 4, 5, 6 and 7, we apply the theorems to a variety of statistical problems and derive the corresponding posterior convergence rates. In Section 8, we extend our discussion from Hölder class to Soblev and Besov function spaces. Finally, a numerical study is presented in Section 9.

2 Notations

Denote $\mathbb{N} = \{1, 2, \ldots\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and $\Omega_j = \{(x_1, \ldots, x_j) : \sum_{i=1}^{j} x_i = 1, x_1, \ldots, x_j \geq 0\}$. Let $\|\boldsymbol{x}\|_p = \{\sum_{i=1}^{d} |x_i|^p\}^{1/p}$ stand for the ℓ_p-norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$; $1 \leq p < \infty$ and $\|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq d} |x_i|$. Similarly, we define $\|f\|_p = \{\int |f(x)|^p dx\}^{1/p}$ and $\|f\|_\infty = \sup_x |f(x)|$ as the L_p-norm of a function f for $1 \leq p \leq \infty$. Define function spaces $L_p = \{f : \|f\|_p < \infty\}$. For a probability measure G, define $\|f\|_{p,G} = \{\int |f(x)|^p dG(x)\}^{1/p}$.

We define δ_x as the point mass probability distribution at point x. Define an indicator function of a set A as $\mathbb{1}\{A\}$.

We define the α-Hölder class C^α as the collection of functions f that has

bounded derivatives up to the order α_0, which is the largest integer strictly smaller than α, and the α_0-th derivative of f satisfies the Hölder condition

$$|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(y)| \leq C|x-y|^{\alpha-\alpha_0} \tag{2.1}$$

for constant $C > 0$ and any x, y in the support of f.

For two k-dimensional vector $\boldsymbol{s}$ and $\boldsymbol{t}$, define $|\boldsymbol{s}| = \sum_{i=1}^k s_i$, $\boldsymbol{s}! = \prod_{i=1}^k s_i!$ and $\boldsymbol{s}^{\boldsymbol{t}} = \prod_{i=1}^k s_i^{t_i}$.

We use $\lesssim$ for inequality up to a constant multiple, where the underlying constant of proportionality is universal or not important for our purposes. If two functions f and g satisfy $f \lesssim g \lesssim f$, we shall write $f \asymp g$.

The Hellinger distance $h(p,q)$ and the Kullback-Leibler (KL) divergence $K(p,q)$ between two densities p and q are commonly used in statistics. They are respectively defined by $h^2(p,q) = \int(\sqrt{p}-\sqrt{q})^2 d\mu$ and $K(p,q) = \int p\log(p/q)d\mu$. Also define the second order KL divergence by $V(p,q) = \int p\log^2(p/q)d\mu$. We define a KL ball around p with radius ϵ as $\mathcal{K}(p,\epsilon) = \{f : K(p,f) \leq \epsilon^2, V(p,f) \leq \epsilon^2\}$.

We use $D(\epsilon, T, d)$ to denote the packing number, which is defined as the maximum cardinality of an ϵ-dispersed subset of T with respect to distance d.

3 General results

3.1 Main theorem

We consider a random variable J taking values in $\mathbb{N}$. For each $J \in \mathbb{N}$, we consider a set of basis functions $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_J)^T$ defined on a measurable space Q. A prior is assigned on J and the coefficients of basis functions $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T$ as follows:

(A1) The prior for J satisfies $\Pi(J > j) \leq A(j)$ and $\Pi(j < J < C_0 j) \geq B(j)$ when j is sufficiently large for some constant $C_0 > 1$. The functions $A(j)$ and $B(j)$ are assumed to be nonnegative and strictly decreasing to 0 when $j \to \infty$.

(A2) Given J, we consider a J-dimensional joint distribution G_1 as the prior for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T$. Assume G_1 satisfies $G_1(\|\boldsymbol{\theta}-\boldsymbol{\theta_0}\|_2 \leq \epsilon) \geq \exp\{-c_2 J\log(1/\epsilon)\}$ for some $\boldsymbol{\theta_0} \in \mathbb{R}^J$, constant $c_2 > 0$ and sufficiently small $\epsilon > 0$.

The priors on J and $\boldsymbol{\theta}$ are allowed to depend on n provided the constants appearing in (A1) and (A2) are free of n. However, to simplify notation, we shall drop the subscript n from Π. In the following, with some abuse of notation, we will use Π as the prior distribution on J and $\boldsymbol{\theta}$ as well as for the induced prior distribution on functions $\boldsymbol{\theta}^T\boldsymbol{\xi}$.

We define two distance metrics d_1 and d_2 on Q satisfying the following conditions:

$$\begin{aligned} d_1(\boldsymbol{\theta_1}^T\boldsymbol{\xi}, \boldsymbol{\theta_2}^T\boldsymbol{\xi}) &\leq \|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_2/a(j), & (3.1)\\ d_2(\boldsymbol{\theta_1}^T\boldsymbol{\xi}, \boldsymbol{\theta_2}^T\boldsymbol{\xi}) &\leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2/b(j) & (3.2) \end{aligned}$$

for some positive functions $a(\cdot)$ and $b(\cdot)$ and every $j \in \mathbb{N}$, $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^j$.

Now we state the main theorem, which can be regarded as a master theorem where the proofs of required conditions for posterior convergence rates for various inference problems are established similarly to Theorem 2.1 of van der Vaart and van Zanten [2008] and Theorem 3.1 of van der Vaart and van Zanten [2009].

Theorem 1. *Let $\epsilon_n \geq \bar{\epsilon}_n$ be two sequence of positive numbers satisfying $\epsilon_n \to 0$ and $n\bar{\epsilon}_n^2 > 1$ as $n \to \infty$. For a function w_0, let there exist two series of positive numbers J_n and M_n, a strictly decreasing, nonnegative function $e(\cdot)$ and a $\boldsymbol{\theta}_{0,j} \in \mathbb{R}^j$ for any $j \in \mathbb{N}$, such that*

$$\begin{aligned} \|\boldsymbol{\theta}_{0,j}\|_\infty &\leq M_n, & (3.3)\\ d_2(w_0, \boldsymbol{\theta}_{0,j}^T\boldsymbol{\xi}) &\leq e(j), & (3.4)\\ \max\{e^{-1}(\bar{\epsilon}_n), n\bar{\epsilon}_n^2\} &\leq J_n, & (3.5)\\ A(J_n) &\lesssim \exp\{-4n\bar{\epsilon}_n^2\}, & (3.6)\\ \log\{1/B(J_n)\} &\lesssim J_n \log(1/\bar{\epsilon}_n), & (3.7)\\ J_n \log J_n + J_n \log\left(\tfrac{a(J_n)M_n}{\epsilon_n}\right) &\lesssim n\epsilon_n^2, & (3.8)\\ \Pi\big(\boldsymbol{\theta} \notin [-M_n, M_n]^j\big) &\lesssim \exp\{-4n\bar{\epsilon}_n^2\}, 1 \leq j \leq J_n, & (3.9) \end{aligned}$$

Let $\mathcal{W}_{J_n,M_n} = \{w = \boldsymbol{\theta}^T\boldsymbol{\xi} : \boldsymbol{\theta} \in \Omega_j, j \leq J_n, \|\boldsymbol{\theta}\|_\infty \leq M_n\}$. Then the following assertions hold:

$$\begin{aligned} \log D(\epsilon_n, \mathcal{W}_{J_n,M_n}, d_1) &\lesssim n\epsilon_n^2, & (3.10)\\ \Pi(W \notin \mathcal{W}_{J_n,M_n}) &\lesssim \exp\{-4n\bar{\epsilon}_n^2\}, & (3.11)\\ -\log\Pi\{w = \boldsymbol{\theta}^T\boldsymbol{\xi} : d_2(w_0, w) \leq 2\bar{\epsilon}_n\} &\lesssim J_n \log\big(\frac{1}{b(J_n)\bar{\epsilon}_n}\big). & (3.12) \end{aligned}$$

Proof. We first verify (3.10), using the definition of packing number, the assumptions on M_n, J_n and (3.1), we obtain:

$$\begin{aligned}
&\log D(\epsilon_n, \mathcal{W}_{J_n,M_n}, d_1) \\
&\leq \log\Big(\sum_{j=1}^{J_n} D(\epsilon_n/a(j), \{\boldsymbol{\theta} \in \Omega_j, \|\boldsymbol{\theta}\|_\infty \leq M_n\}, \|\cdot\|_2)\Big) \\
&\leq \log\Big[J_n\big\{\frac{a(J_n)\sqrt{J_n}M_n}{\epsilon_n}\big\}^{J_n}\Big] \\
&\lesssim J_n \log\Big(\frac{a(J_n)\sqrt{J_n}M_n}{\epsilon_n}\Big) \leq n\epsilon_n^2 \qquad (3.13)
\end{aligned}$$

Next, to verify (3.11), observe the following:

$$\begin{aligned}
\Pi(w \notin \mathcal{W}_{J_n,M_n}) &\leq \Pi(J > J_n) + \sum_{j=1}^{J_n} \Pi(\boldsymbol{\theta} \notin [-M_n, M_n]^j)\Pi_n(J = j) \\
&\leq A(J_n) + \exp\{-4n\bar{\epsilon}_n^2\} \\
&\lesssim \exp\{-4n\bar{\epsilon}_n^2\}. \qquad (3.14)
\end{aligned}$$

For (3.12), using (3.3) and (3.4), since $d_2(w_0, \boldsymbol{\theta}_{0,j}^T\boldsymbol{\xi}) \leq e(j) \leq \bar{\epsilon}_n$ for all $j \geq J_n$, we have

$$\begin{aligned}
\Pi\{w : d_2(w_0, \boldsymbol{\theta}^T\boldsymbol{\xi}) \leq 2\bar{\epsilon}_n\} &\geq \Pi(J_n \leq J \leq C_1 J_n) G_1(\|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|_2 \leq b(J_n)\bar{\epsilon}_n) \\
&\geq B(J_n)\exp\{-c_2 J_n \log\big(\frac{1}{b(J_n)\bar{\epsilon}_n}\big)\} \qquad (3.15)
\end{aligned}$$

for some constants $C_1 > 1$ and $c_2 > 0$. By taking a negative logarithm on both sides and using (3.7), it follows that (3.9) holds. Hence the proof is complete. □

Remark 1. The choice of M_n is closely related to the prior distribution G_1 on $\boldsymbol{\theta}$. For example, if we restrict the prior on a bounded set, then M_n can be chosen as a large constant such that the left hand side of (3.9) equals to 0. If we assume exponential decay condition $G_1(\boldsymbol{\theta} \notin [-M, M]^j) \leq \exp\{-c_3 j M^{c_4}\}$ for some constants $c_3, c_4 > 0$, all $j \in \mathbb{N}$ and sufficiently large M, then M_n can be chosen as a polynomial in n.

Remark 2. Conditions (3.4), (3.5), (3.6) and (3.7) in Theorem 1 require a sufficiently large J_n in order to have sufficiently good approximation to w_0. In

other words, J_n controls the bias of the model. Meanwhile, Conditions (3.8) and (3.9) state that J_n should not be too large if the complexity of the model is to be controlled. When studying Bayesian asymptotic properties, a balance between bias and complexity needs to be established to obtain the optimal posterior convergence.

Next, we give some examples to illustrate the use of Theorem 1.

Example 1. *Fourier trigonometric series*
Choose the basis $\{\cos jx, \sin jx, j \in \mathbb{N}\}$, then for a function $w_0 \in C^\alpha$, we have $e(J) = C(\log J)/J^\alpha$ for some constant C and d_2 as the supremum distance [Jackson, 1930]. Therefore, we can choose $J_n = n^{1/(2\alpha+1)}(\log n)^{2/(2\alpha+1)}$, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}(\log n)^{1/(2\alpha+1)}$. The functions $A(\cdot)$ and $B(\cdot)$ in the prior can be chosen as exponential functions $A(x) = \exp\{-c_1 x\}$ and $B(x) = \exp\{-c_2 x \log x\}$ for some positive constants c_1 and c_2. Then the rate ϵ_n is $n^{-\alpha/(2\alpha+1)}(\log n)^{1/(2\alpha+1)+1/2}$.

Example 2. *Bernstein polynomials*
We consider the Bernstein polynomial prior proposed by Petrone [1999b,a]. Consider a continuously differentiable density function w_0 with bounded second derivative, the approximation property of Bernstein polynomials to w_0 is $e(J) = C/J$ for some universal constant C and d_2 as the supremum distance [Lorenz, 1953]. We can choose $J_n = n^{1/3}$, $\bar{\epsilon}_n = n^{-1/3}$ and again same choices for $A(\cdot)$ and $B(\cdot)$ as in Example 1. The rate ϵ_n is $n^{-1/3}(\log n)^{1/2}$, which has the same polynomial power as given in Ghosal [2001].

Example 3. *Polynomial basis*
Consider the orthogonal polynomials as the approximation tool for $w_0 \in C^\alpha([0,1])$, we have $e(J) = CJ^{-\alpha}$ for some universal constant C and d_2 as the L_2-distance (e.g., Theorem 6.1 of Hesthaven et al. [2007]). Then we can choose $J_n = n^{1/(2\alpha+1)}$, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}$ and the same $A(\cdot)$, $B(\cdot)$ as in Example 1. The rate ϵ_n will be $\bar{\epsilon}_n$ multiplied by some power of $\log n$, where the power depends on the statistical problem.

Note that although the approximation is given for L_2-distance instead of supremum distance, the adaptive rate can still be obtained under mild conditions, see Remarks 5, 7, 8, 9, 10 and 12.

Example 4. *B-splines*
Choose B-spline as the basis functions, then for $w_0 \in C^\alpha([0,1])$, we have $e(J) \asymp J^{-\alpha}$ for d_2 as the supremum distance. Then we can choose $J_n = n^{1/(2\alpha+1)}$, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}$ and ϵ_n as $n^{-\alpha/(2\alpha+1)}$ multiplied by some power of $\log n$, where

the power depends on the statistical problem. More details are given in Section 3.2.

Example 5. *Wavelets*
We consider a multiresolution wavelet:

$$\sum_{k\in\mathbb{Z}} \alpha_{0k}\phi_{0k}(x) + \sum_{j=0}^{j_1}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}(x), \tag{3.16}$$

where ϕ are father wavelets, ψ are mother wavelets and $j_1 \in \mathbb{N}$. We put priors on j_1 and wavelet coefficients α and β. It has been show that, for $w_0 \in C^\alpha$, the approximation error is $e(j_1) = 2^{-j_1\alpha}$ for d_2 as the supremum distance [e.g. Giné and Nickl [2009]]. Meanwhile, if w_0 has a compact support, the number of nonzero coefficients is $O(2^{j_1})$. Hence we apply Theorem 1 for $J = 2^{j_1}$ and choose $J_n = n^{1/(2\alpha+1)}$, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}$ and $A(\cdot)$, $B(\cdot)$ as in Example 1. The resulting rate ϵ_n is $n^{-\alpha/(2\alpha+1)}(\log n)^{1/2}$. This coincides with the adaptation results for white noise models in Lian [2011] and for density estimation and regression models in Rivoirard and Rousseau [2012].

3.2 B-splines

We consider the B-spline model as described in de Boor [2001]. For a given compact interval on the real line, we divide it into K equally spaced subintervals and then define a class of B-spline basis functions that are q times differentiable. Then it can be shown that these spline functions form a J-dimensional linear space, where J is defined as $J = q+K-1$. In our study, these B-spline functions are used to approximate the underlying data generation scheme (e.g. true densities). The approximation ability of the B-spline functions is determined by the smoothness level of the data α and the number of spline basis functions J under the condition that $q \geq \alpha$. Denote $\boldsymbol{B}$ as the column vector of B-spline basis functions. In the following discussion, we assume q is chosen large enough such that $q \geq \alpha$ holds for every statistical problem we are interested in.

Lemma 1. *(1) For any function $f \in C^\alpha([0,1])$, $0 < \alpha \leq q$, there exists $\boldsymbol{\theta} \in \mathbb{R}^J$ and a constant $C > 0$ such that*

$$\|f - \boldsymbol{\theta}^T\boldsymbol{B}\|_\infty \leq CJ^{-\alpha}\|f^{(\alpha)}\|_\infty. \tag{3.17}$$

(2) Further, if $f > 0$ and $J > J_0$, where J_0 is a sufficiently large constant that only depends on f and q, then every element of $\boldsymbol{\theta}$ can be chosen to be positive.

Proof. The first part is a well-known spline approximation result [de Boor, 2001]. For the second assertion, assume $f \geq \epsilon$ for some $\epsilon > 0$, use the result in de Boor [2001, Chap. 6], for each θ_i, there exists a universal constant C_1 that depends only on q and f, such that $|\theta_i - c| \leq C_1 \sup_{x \in [t_{i+1}, t_{i+q-1}]} |f(x) - c|$ for any choice of constant c. Here t_{i+1} and t_{i+q-1} are $(i+1)$-th and $(i+q-1)$-th knots. Choose $c = \inf_{x \in [t_{i+1}, t_{i+q-1}]} f(x) \geq \epsilon$, then the right hand side is bounded by $C_1(q/J)^\alpha$. Choose $J > q(C_1/\epsilon)^{1/\alpha}$, we have $\theta_i > c - C_1(q/J)^\alpha \geq 0$. □

Remark 3. In part (2), the condition $f > 0$ is crucial. If we approximate a nonnegative function f using nonnegative coefficients $\boldsymbol{\theta}$, then the approximation error is only $O(J^{-1})$ [de Boor and Daniel, 1974], which does not adapt to the smoothness level beyond 1.

For $w_0 \in C^\alpha$, we have $e(J) = J^{-\alpha}$ for d_2 chosen as the supremum distance. This leads to the following choice of priors for J and the coefficients of basis functions $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T$:

(B1) The prior for J satisfy $\Pi(J > j) \leq \exp\{-c_0 j \log^{t_1} j\}$ and $\Pi_n(j < J < C_0 j) \geq \exp\{-c_1 j \log^{t_2} j\}$ for some constants $c_0, c_1 > 0$, $0 \leq t_1 \leq t_2 \leq 1$ and $C_0 > 1$ when j is sufficiently large.

(B2) Given J, we consider a J-dimensional joint distribution G_2 as the prior for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T$. Assume G_2 satisfies $G_2(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|_2 \leq \epsilon) \geq \exp\{-c_2 J \log(1/\epsilon)\}$ for some $\boldsymbol{\theta_0} \in \mathbb{R}^J$, constant $c_2 > 0$ and small ϵ.

(B3) In particular, if $w_0 > 0$, then we allow the prior for $\boldsymbol{\theta}$ being constructed on $(0, \infty)^J$ satisfying $\Pi(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta_0}\|_2 \leq \epsilon) \geq \exp\{-c_2 J \log(1/\epsilon)\}$.

Remark 4. We give some examples of such priors. Geometric, Poisson and negative binomial distributions satisfy Condition (B1). Normal, gamma, exponential, Dirichlet distributions satisfy Condition (B2); see Lemma 6.1 of Ghosal et al. [2000] for the last conclusion.

Using the relation $\|\boldsymbol{\theta_1}^T B - \boldsymbol{\theta_2}^T B\|_1 \lesssim \|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_2/\sqrt{J}$ and $\|\boldsymbol{\theta_1}^T B - \boldsymbol{\theta_2}^T B\|_\infty \lesssim \|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_2/\sqrt{J}$ for $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^J$, we get the following conclusion from Theorem 1.

Theorem 2. *Let $\epsilon_n \geq \bar{\epsilon}_n$ be two sequence of positive numbers satisfying $\epsilon_n \to 0$ and $n\bar{\epsilon}_n^2 > 1$ as $n \to \infty$. If there exist J_n and M_n satisfying the following*

conditions:

$$\max\{\bar{\epsilon}_n^{-1/\alpha}, n\bar{\epsilon}_n^2\} \leq J_n, \tag{3.18}$$

$$c_1 \log^{t_1} J_n \leq \log(1/\bar{\epsilon}_n), \tag{3.19}$$

$$J_n \log\big(\tfrac{\sqrt{J_n}M_n}{\epsilon_n}\big) \leq n\epsilon_n^2, \tag{3.20}$$

$$\Pi_n\big(\boldsymbol{\theta} \notin [-M_n, M_n]^J\big) \lesssim \exp\{-4n\bar{\epsilon}_n^2\}, 1 \leq J \leq J_n, \tag{3.21}$$

Let $\mathcal{W}_{J_n,M_n} = \{w = \boldsymbol{\theta}^T\boldsymbol{B} : \boldsymbol{\theta} \in \mathbb{R}^J, J \leq J_n, \|\boldsymbol{\theta}\|_\infty \leq M_n\}$*, then the following assertions hold:*

$$\log D(\epsilon_n, \mathcal{W}_{J_n,M_n}, \|\cdot\|_1) \lesssim n\epsilon_n^2, \tag{3.22}$$

$$\Pi(W \notin \mathcal{W}_{J_n,M_n}) \lesssim \exp\{-4n\bar{\epsilon}_n^2\}, \tag{3.23}$$

$$-\log \Pi\{\|w_0 - \boldsymbol{\theta}^T\boldsymbol{B}\|_\infty \leq 2\bar{\epsilon}_n\} \lesssim J_n \log\big(\frac{1}{\sqrt{J_n}\bar{\epsilon}_n}\big). \tag{3.24}$$

4 Gaussian white noise model

We consider a Gaussian white noise model

$$dX(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t),\ 0 \leq t \leq 1, \tag{4.1}$$

where $X(t)$ is the observed signal, $f(t)$ is the unknown signal and $W(t)$ is a standard Wiener process. Let ϕ_i, $i = 1, 2, \ldots$ be an orthonormal basis in $L_2[0,1]$. Assume $f \in L_2[0,1]$, then this problem can be transformed into the estimation of the mean $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ for an infinite-dimensional normal distribution as follows:

$$X_i = \theta_i + \frac{\epsilon_i}{\sqrt{n}},\ i = 1, 2, \ldots. \tag{4.2}$$

Here X_i's are independent observations, ϵ_i's are i.i.d white noise variables that follows a normal distribution $N(0,1)$. Asymptotic results have been obtained in frequentist studies. For example, in Pinsker [1980], the minimax convergence rate is shown to be $n^{-\alpha/(2\alpha+1)}$ for L_2-class with respect to quadratic risk. In a Bayesian study, Belitser and Ghosal [2003] obtained the same posterior convergence rate for a discrete collection of smoothness parameters α. Babenko and Belitser [2010] considered putting a prior on the oracle J, which is defined as the best cut-off $\theta_i = 0$ for all $i > J$ such that the risk of

$\{X_i \mathbb{1}(i \le J) : i \in \mathbb{N}\}$ is minimized. They showed that such an oracle estimator is minimax for a general smoothness class and hence obtained adaptation results as well.

Assume the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ belongs to a Soblev type class $\mathcal{S}(\alpha) = \{\boldsymbol{\theta} : \sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \le Q\}$ for some $\alpha, Q > 0$. Here α can be viewed as a smoothness parameter. We consider the oracle $J > 0$ such that $\theta_i = 0$ for all $i > J$. The approximation error is

$$e(J) \asymp \frac{J}{n} + \sum_{i=J+1}^{\infty} \theta_i^2 \le \frac{J}{n} + \sum_{i=J+1}^{\infty} \frac{\theta_i^2 i^{2\alpha}}{n^{2\alpha}} \le \frac{J}{n} + \frac{Q}{J^{2\alpha}}. \tag{4.3}$$

Then by constructing appropriate priors on $\boldsymbol{\theta}$ and J as in Section 3.1, using Theorem 1 for $J_n = n^{1/(2\alpha+1)}$, $\bar{\epsilon}_n$, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}$ and d_2 as ℓ_2-distance, we obtain the following theorem:

Theorem 3. *Suppose the true mean $\boldsymbol{\theta_0} \in \mathcal{S}(\alpha)$ for some $0 < \alpha \le q$. The prior is constructed as in Section 3.1, then the posterior distribution of $\boldsymbol{\theta}$ converges at rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{1/2}$ with respect to ℓ_2-distance.*

5 Density estimation

5.1 Density on a known compact interval

In this section, we apply the general results to density estimation problems. In the frequentist study, optimal rate of convergence $n^{-\alpha/(2\alpha+1)}$ has been obtained for the maximum likelihood estimators in Stone [1990]. A Bayesian log-spline model has been studied in Ghosal et al. [2000], where the optimal posterior convergence rate is shown to be $n^{-\alpha/(2\alpha+1)}$. When α is unknown, the rate $n^{-\alpha/(2\alpha+1)}$ up to an additional logarithmic factor is established in Ghosal et al. [2008].

We first consider estimating a density function f_0 that is defined on a unit interval $[0, 1]$. A Bayesian estimator of f can be constructed by using B-spline functions through a nonnegative, monotonic link function g, i.e., $p_{\boldsymbol{\theta}} = g(\sum_{l=1}^{J} \theta_l B_l)/\int_0^1 g(\sum_{l=1}^{J} \theta_l B_l)$ for $\boldsymbol{\theta} \in \mathbb{R}^J$ and J is given a prior on $\mathbb{N}$. We can choose g as the identity function, and restrict the prior for $\boldsymbol{\theta}$ on $(0, \infty)^J$. If we choose g as the exponential function, then it gives the log-spline model. If g is a polynomial, $p_{\boldsymbol{\theta}}$ is a rational function of $\boldsymbol{\theta}$. By using identity link function, it is possible to avoid the normalization altogether; see Section 5.3.

The theorem of the posterior convergence rate is stated as follows:

Theorem 4. *Consider n independent, identically distributed observations $X_1, \ldots, X_n$ from the true density $f_0 \in C^\alpha[0,1]$ and $\alpha \leq q$. Suppose that f_0 is positive. If the prior is constructed as in Section 3.2, then the posterior distribution converges at rate $\epsilon_n = n^{-\alpha(2\alpha+1)}(\log n)^{1/2}$ with respect to the Hellinger or the L_2-distance.*

Proof. From the spline approximation theory, given J, there exists $\boldsymbol{\theta_0} \in \mathbb{R}^J$ such that $\|\boldsymbol{\theta_0}^T \boldsymbol{B} - g^{-1}(f_0)\|_\infty \lesssim J^{-\alpha}$. This gives $\|f_0 - g(\boldsymbol{\theta_0}^T \boldsymbol{B})\|_\infty \lesssim J^{-\alpha}$. By integration, we have $|1 - \int_0^1 g(\boldsymbol{\theta_0}^T \boldsymbol{B})(x)dx| \lesssim J^{-\alpha}$. Hence $\|f_0 - p_{\boldsymbol{\theta_0}}\|_\infty \lesssim J^{-\alpha}$ and $h(f_0, p_{\boldsymbol{\theta_0}}) \lesssim J^{-\alpha}$. Using Lemma 8 of Ghosal and van der Vaart [2007a], we have

$$K(f_0, p_{\boldsymbol{\theta_0}}) \leq 2h^2(f_0, p_{\boldsymbol{\theta_0}}) \Big\| \frac{f_0}{p_{\boldsymbol{\theta_0}}} \Big\|_\infty \lesssim J^{-2\alpha},$$
$$V(f_0, p_{\boldsymbol{\theta_0}}) \lesssim h^2(f_0, p_{\boldsymbol{\theta_0}}) \Big(1 + \Big\| \frac{f_0}{p_{\boldsymbol{\theta_0}}} \Big\|_\infty \Big)^2 \lesssim J^{-2\alpha}. \tag{5.1}$$

Now we use Lemma 7.4 in Ghosal et al. [2008], which states that the Helligner distance on $p_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \mathbb{R}^J$ is equivalent to the L_2-distance on $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^J$ divided by $\sqrt{J}$:

$$\inf_{x,\boldsymbol{\theta}} \Big(\frac{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2}{J} \wedge 1 \Big) \lesssim h^2(p_{\boldsymbol{\theta}_1}, p_{\boldsymbol{\theta}_2}) \lesssim \sup_{x,\boldsymbol{\theta}} \Big(\frac{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2}{J} \wedge 1 \Big). \tag{5.2}$$

Hence we can apply Lemma 7.6 in that paper and calculate the prior probability on an L_2-ball around $\boldsymbol{\theta_0}$ instead of a KL-ball around f_0. We apply Theorem 2 for $J_n \asymp n^{1/(2\alpha+1)}$, M_n as a constant satisfying $M_n > \max\{g^{-1}(f_0), f_0\}$, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}$ and $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{1/2}$. Note that (3.24) and (5.1) together imply $\Pi(\mathcal{K}(f_0, \epsilon_n)) \geq \exp\{-cn\epsilon_n^2\}$ for some positive constant c. Then we apply Theorem 2.1 of Ghosal et al. [2000], the proof is complete. □

Remark 5. If we assume $\alpha \geq 1$, namely, f_0 is Lipschitz continuous, then the proof above can proceed using L_2-approximation only. From $\|\boldsymbol{\theta_0}^T \boldsymbol{B} - g^{-1}(f_0)\|_2 \lesssim J^{-\alpha}$, we have $\|f_0 - g(\boldsymbol{\theta_0}^T \boldsymbol{B})\|_1 \leq \|f_0 - g(\boldsymbol{\theta_0}^T \boldsymbol{B})\|_2 \lesssim J^{-\alpha}$. Hence $\|f_0 - p_{\boldsymbol{\theta_0}}\|_2 \leq 2\|f_0 - g(\boldsymbol{\theta_0}^T \boldsymbol{B})\|_2 \lesssim J^{-\alpha}$. Then $h(f_0, p_{\boldsymbol{\theta_0}}) = \|\sqrt{f_0} - \sqrt{p_{\boldsymbol{\theta_0}}}\|_2 \lesssim \|f_0 - p_{\boldsymbol{\theta_0}}\|_2 \lesssim J^{-\alpha}$ since f_0 is bounded below. The KL divergences can be bounded similarly. This implies that we can use other basis expansions such as orthogonal polynomials instead of B-spline functions as long as we have L_2-approximation results.

5.2 Density with unbounded support

Let $f_0 \in C^\alpha(\mathbb{R})$. Consider a fixed monotonic link function $\Psi : \mathbb{R} \to [0,1]$ to change the domain of the function to $[0,1]$. Suppose that there is an interval $[a,b]$ such that f_0 is bounded away from 0 on it. Consider a pseudo-metric $d(f_1,f_2) = \int_a^b |f_1(y) - f_2(y)|dy$. By constructing a prior on f through the representation $f(y) = g\{\boldsymbol{\theta}^T \boldsymbol{B}(\Psi(y))\}$ and arguing as in Theorem 4, we obtain the same posterior convergence rates with respect to d. The same method also applies for half open intervals.

5.3 Computation

Consider the setting of Subsection 5.1. It is well known that [see Schumaker, 2007, Chap. 4]

$$\int_0^1 B_i(x)dx = \begin{cases} \frac{i}{q(J-q+1)} & i = 1,\ldots,(q-1); \\ \frac{1}{J-q+1} & i = q,\ldots,(J-q+1); \\ \frac{J-i+1}{q(J-q+1)} & i = (J-q+2),\ldots,J. \end{cases} \tag{5.3}$$

Define scaled B-spline basis functions $B_k^* = B_i / \int_0^1 B_i$, $i = 1,\ldots,J$ so that $\int_0^1 B_i^*(x)dx = 1$, $i = 1,\ldots,J$.

We restrict the coefficients $\boldsymbol{\theta}$ to satisfy $\sum_{k=1}^J \theta_k = 1$ and form a density $f = \sum_{k=1}^J \theta_k B_k^*$. We put a Dirichlet prior on $\boldsymbol{\theta} \sim \mathrm{Dir}(a_1, a_2, \ldots, a_J)$ given fixed J. Finally, we assign a prior Π on J. Thus a prior on the density f is induced. Given the observations $X_1,\ldots,X_n$ and a fixed dimension J, the posterior density of $\boldsymbol{\theta}$ is a mixture of Dirichlet distribution:

$$\begin{aligned} p(\boldsymbol{\theta}|X_1,\ldots,X_n,J) &\propto \prod_{k=1}^J \theta_k^{a_k-1} \prod_{i=1}^n \Big\{ \sum_{k=1}^J \theta_k B_k^*(X_i) \Big\} \\ &= \sum_{i_1=1}^J \cdots \sum_{i_n=1}^J \prod_{k=1}^J \theta_k^{a_k-1} \prod_{s=1}^n \theta_{i_s} B_{i_s}^*(X_s). \end{aligned} \tag{5.4}$$

Thus, the posterior mean of f at a point x is

$$\frac{\sum_{j=1}^\infty \Pi(J) \sum_{i_0=1}^j \sum_{i_1=1}^j \cdots \sum_{i_n=1}^j \int_{\boldsymbol{\theta}\in\Omega_j} \prod_{k=1}^j \theta_k^{a_k-1} \prod_{s=0}^n \theta_{i_s} B_{i_s}^*(x_s) d\boldsymbol{\theta}}{\sum_{j=1}^\infty \Pi(J) \sum_{i_1=1}^j \cdots \sum_{i_n=1}^j \int_{\boldsymbol{\theta}\in\Omega_j} \prod_{k=1}^j \theta_k^{a_k-1} \prod_{s=1}^n \theta_{i_s} B_{i_s}^*(x_s) d\boldsymbol{\theta}}, \tag{5.5}$$

where X_0 stands for x. Define $I_{k,j,0} = \sum_{s=0}^{j} \mathbb{1}\{i_s = k\}$ and $I_{k,j,1} = \sum_{s=1}^{j} \mathbb{1}\{i_s = k\}$. The quantity above can be simplified into

$$\frac{\sum_{j=1}^{\infty} \Pi(J) \sum_{i_0=1}^{j} \sum_{i_1=1}^{j} \cdots \sum_{i_n=1}^{j} \prod_{k=1}^{j} \Gamma(a_k + I_{k,j,0}) \prod_{s=0}^{n} B_{i_s}^*(x_s) / \Gamma\Big(\sum_{i=1}^{j} a_i + n + 1\Big)}{\sum_{j=1}^{\infty} \Pi(J) \sum_{i_1=1}^{j} \cdots \sum_{i_n=1}^{j} \prod_{k=1}^{j} \Gamma(a_k + I_{k,j,1}) \prod_{s=0}^{n} B_{i_s}^*(x_s) / \Gamma\Big(\sum_{i=1}^{j} a_i + n\Big)}. \quad (5.6)$$

A basis function takes nonzero values only at q intervals, so the calculation involves a multiple of q^{n+1} loops. More details are given in Section 9. Similar expressions can be obtained for posterior moments, in particular, the posterior variance.

Remark 6. Note that the adaptive Bayesian estimator has a connection with histogram and kernel methods. Namely, if $q = 1$, the sums over indices $i_1, \ldots, i_n$ in (5.6) will vanish. Hence it gives a histogram estimate of densities, which is similar with the Bayes estimator in Gasparini [1996]. Although doing so brings easy computation, it cannot adapt to the smoothness level greater than 1. Our estimator can also be viewed as a kernel estimator, where the kernel is induced by a discrete parameter and the kernel takes positive values only in a finite interval. When q and J are chosen larger, the kernel becomes more flat.

6 Whittle estimation of spectral density

6.1 Posterior convergence rates

Consider a second order stationary time series $\{X_t, t \in \mathbb{Z}\}$ with mean 0 and autocovariance function $\gamma_r = \mathrm{E}(X_t X_{t+r})$. Assume $\sum_r |\gamma_r| < \infty$. Define the spectral density of $\{X_t\}$ by $f(\lambda) = (2\pi)^{-1} \sum_{r=-\infty}^{\infty} \gamma_r e^{-ir\pi\lambda}$ on $[0, 1]$. Let $I_n(\lambda) = (2\pi n)^{-1} |\sum_{t=1}^{n} X_t e^{-it\pi\lambda}|^2$ be the periodogram. Instead of using the true likelihood, which is complicated even for a Gaussian time series, Whittle [1957] proposed using an approximate likelihood. Let $\nu = \lfloor n/2 \rfloor$ and $\omega_j = 2j/n$, $j = 1, \ldots, \nu$. The Whittle likelihood is formed by pretending that $U_j = I_n(\omega_j)$, $j = 1, \ldots, \nu$ with means $f(\omega_j)$ are exponentially distributed. The advantage of using the Whittle likelihood is that it is an explicit function of the spectral density f, rather than being expressed in terms of infinitely many autocorrelation coefficients.

A nonparametric Bayesian method has been used by Choudhuri et al. [2004a] where the prior is constructed via Bernstein polynomials. They also gave a posterior consistency result. Posterior convergence rates can be obtained by a contiguity result as established in Choudhuri et al. [2004b]. Their results state that for a Gaussian time series, the distributions of $(U_1, \ldots, U_\nu)$ given by the exact and the Whittle likelihood are contiguous. As long as asymptotic convergence results are concerned, this allows us to work under the pretending assumption that $U_1, \ldots, U_\nu$ are actually independent and exponentially distributed.

Assume that the true spectral density $f_0 \in C^\alpha[0,1]$ is positive throughout. A prior is constructed on f using a monotonic link function g. For example, if g is the identity function, then the prior for $\boldsymbol{\theta}$ can be constructed on $(0,\infty)^J$ and $f = f_{\boldsymbol{\theta}}$ is linear in $\boldsymbol{\theta}$. If $g(x) = \log x$, then the prior for $\boldsymbol{\theta}$ is constructed on $\mathbb{R}^J$. The posterior rates will be the same in both cases because the sieve in Theorem 2 will not be affected as long as M_n is bounded by a polynomial of n. Define a discretized L_2-distance as follows:

$$d_n^2(f_1, f_2) = \nu^{-1} \sum_{i=1}^{\nu} \{f_1(2i/n) - f_2(2i/n)\}^2. \tag{6.1}$$

Then we have the following theorem.

Theorem 5. *Assume that the true spectral density $f_0 \in C^\alpha([0,1])$, takes value in a bounded interval $[m, M]$ and $\alpha \le q$. Suppose that the prior is constructed as in Section 3.2. Then the posterior distribution converges at rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{1/2}$ with respect to d_n. If $\alpha \ge 1$, then the result also holds for the L_2-distance.*

Proof. Use results in Section 7.3 of Ghosal and van der Vaart [2007b],

$$\max\{\nu^{-1} \sum_{i=1}^{\nu} K(P_{f_0,i}, P_{f,i}), \nu^{-1} \sum_{i=1}^{\nu} V(P_{f_0,i}, P_{f,i})\} \lesssim d_n^2(f_0, f) \lesssim \|f_0 - f\|_\infty^2 \tag{6.2}$$

Then use the same arguments as in Theorem 4, apply Theorem 4 of Ghosal and van der Vaart [2007b], the conclusion holds. If $\alpha \ge 1$, then f_0 is Lipschitz continuous. Hence $\|f_1 - f_2\|_2 \lesssim d_n(f_1, f_2) + (L+M)/n$ for any density functions $f_1, f_2 \le M$ having Lipschitz constant L. Therefore d_n can be substituted by L_2. □

Remark 7. When $\alpha \ge 1$, observe $|f_1(2j/n) - f_2(2j/n)| \le |f_1(x) - f_2(x)| + 4L/n$ for $x \in [2(j-1)/n, 2j/n)$ and $j = 1, \ldots, \nu$. By integrating and taking squares

of both sides, we have $d_n^2(f_1, f_2) \lesssim \|f_1 - f_2\|_2^2 + L^2/n^2 + ML/n$. This implies that we only need L_2-approximation to get adaptive rates. Hence other basis expansion techniques that has an L_2-approximation property can also be used to construct the Bayesian estimator for this model.

6.2 Computation

Choose the link function $g(x) = x^{-1}$, the likelihood is

$$\frac{1}{\prod_{s=1}^{\nu} f(\omega_s)} \exp\{-\sum_{s=1}^{\nu} U_s / f(\omega_s)\}. \tag{6.3}$$

We consider independent gamma prior distributions on $\boldsymbol{\theta}$: $\theta_i \overset{\text{ind}}{\sim} \text{Gamma}(a_i, b_i)$ for some positive numbers a_i and b_i. The posterior mean of $g(f_0)$ is given by

$$\frac{\sum_{j=1}^{\infty} \Pi(j) \sum_{i_0=1}^{j} \cdots \sum_{i_\nu=1}^{j} \int_{\boldsymbol{\theta}>0} \prod_{k=1}^{j} \frac{\theta_k^{a_k-1} b_k^{a_k}}{\Gamma(a_k)} \prod_{s=0}^{\nu} \theta_{i_s} B_{i_s}(\omega_s) \exp\{-\sum_{s=1}^{\nu} U_s \sum_{i=1}^{j} \theta_i B_i(\omega_s) - \sum_{i=1}^{j} b_i \theta_i\} d\boldsymbol{\theta}}{\sum_{j=1}^{\infty} \Pi(j) \sum_{i_1=1}^{j} \cdots \sum_{i_\nu=1}^{j} \int_{\boldsymbol{\theta}>0} \prod_{k=1}^{j} \frac{\theta_k^{a_k-1} b_k^{a_k}}{\Gamma(a_k)} \prod_{s=1}^{\nu} \theta_{i_s} B_{i_s}(\omega_s) \exp\{-\sum_{s=1}^{\nu} U_s \sum_{i=1}^{j} \theta_i B_i(\omega_s) - \sum_{i=1}^{j} b_i \theta_i\} d\boldsymbol{\theta}}.$$

Define $I_{k,j,0} = \sum_{s=0}^{j} \mathbb{1}\{i_s = k\}$ and $I_{k,j,1} = \sum_{s=1}^{j} \mathbb{1}\{i_s = k\}$. The quantity above can be simplified into

$$\frac{\sum_{j=1}^{\infty} \Pi(j) \sum_{i_0=1}^{j} \cdots \sum_{i_\nu=1}^{j} \prod_{s=0}^{\nu} B_{i_s}(\omega_w) \prod_{k=1}^{j} \frac{b_k^{a_k}}{\{b_k + \sum_{s=1}^{\nu} U_s B_k(\omega_s)\}^{I_{k,j,0}+a_k}} \frac{\Gamma(I_{k,j,0}+a_k)}{\Gamma(a_k)}}{\sum_{j=1}^{\infty} \Pi(j) \sum_{i_1=1}^{j} \cdots \sum_{i_\nu=1}^{j} \prod_{s=1}^{\nu} B_{i_s}(\omega_w) \prod_{k=1}^{j} \frac{b_k^{a_k}}{\{b_k + \sum_{s=1}^{\nu} U_s B_k(\omega_s)\}^{I_{k,j,1}+a_k}} \frac{\Gamma(I_{k,j,1}+a_k)}{\Gamma(a_k)}} \tag{6.4}$$

7 Nonparametric regression

7.1 Regression with Gaussian errors

We consider a nonparametric regression model with additive error $X_i = f(Z_i) + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and σ is an unknown parameter. The covariates $Z_1, \ldots, Z_n$ can be either fixed or random. We first treat the fixed covariates

case. We will use the same spline-based prior on f as in Section 3. Also, we put a prior distribution G_3 on σ that satisfies a tail condition for sufficiently small $\sigma > 0$ and some positive number c:

$$G_3(\sigma) \lesssim \exp\{-c\exp(1/\sigma)\}. \tag{7.1}$$

In practice, this condition is automatically satisfied if σ has a confirmed lower bound. Define a empirical measure $\mathbb{P}_n^Z = n^{-1}\sum_{i=1}^n \delta_{Z_i}$, the covariance matrix $\Sigma_{n,J}$, whose (k,l)-th element is $\int B_k B_l d\mathbb{P}_n^Z$ for $k, l \le J$ and $\|\cdot\|_{2,n}$ as the norm on $L_2(\mathbb{P}_n^Z)$.

Denote the true value of the regression function as f_0. We need the following assumptions on f_0 and $Z_1, \cdots, Z_n$:

(D1) Smoothness: The function f_0 is assumed to be α-Hölder, where $0 < \alpha \le q$.

(D2) Separation property:

$$\boldsymbol{\theta}^T \Sigma_{n,J} \boldsymbol{\theta} \lesssim J^{-1}\|\boldsymbol{\theta}\|^2,\ \boldsymbol{\theta} \in \mathbb{R}^J. \tag{7.2}$$

When the domain of the density is $\mathbb{R}$, define a pseudo-metric $\|\cdot\|_{2,n}^*$ as $\left(\|f_1 - f_2\|_{2,n}^*\right)^2 = \int_a^b |f_1 - f_2|^2 d\mathbb{P}_n^Z$ similarly as in Section 5.2.

Theorem 6. *Suppose the true regression function f_0 satisfy conditions (D1) and (D2), the covariates $Z_1, \ldots, Z_n$ are fixed and a prior is constructed as in Section 3.2 and* (7.1).
(1) If f_0 is defined on $[0,1]$, then the posterior converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}\log n$ relative to $\|\cdot\|_{2,n}$.
(2) If the support of f_0 is unbounded, e.g., the real line, then the posterior converges at the same rate with respect to $\|\cdot\|_{2,n}^*$.

Proof. We first consider part (1). Define $P_{f,i}$ as the normal measure with mean $f(Z_i)$ and variance σ^2. Results in Birgé [2006] imply that the likelihood ratio test for f_0 versus another f satisfy conditions in Lemma 2 of Ghosal and van der Vaart [2007b] with respect to $\|\cdot\|_{2,n}$; see the latter paper for details. This implies that we can work on $\|\cdot\|_{2,n}$ instead of Hellinger distance. Using the arguments in Section 7.2 of Ghosal and van der Vaart [2007b], we get

$$\max\left\{n^{-1}\sum_{i=1}^n K(P_{f_0,i}, P_{f,i}), n^{-1}\sum_{i=1}^n V(P_{f_0,i}, P_{f,i})\right\} \le \|f_0 - f\|_n^2/\sigma^2. \tag{7.3}$$

Define a sieve by $\mathcal{W}_{J_n,M_n} \bigcap \{\sigma \geq 2\alpha/\log n\}$. Using the fact that $\|f_0 - f\|_{2,n}^2 \lesssim \|f_0 - f\|_\infty^2$, there exists $\boldsymbol{\theta_0} \in \mathbb{R}^J, \|\boldsymbol{\theta_0}\|_\infty < \infty$ such that

$$\max\left\{n^{-1}\sum_{i=1}^n K(P_{f_0,i}, P_{f,i}), n^{-1}\sum_{i=1}^n V(P_{f_0,i}, P_{f,i})\right\} \lesssim J^{-2\alpha}(\log n)^2. \tag{7.4}$$

Under Condition (D2), for all $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^J$:

$$\|f_{\boldsymbol{\theta_1}} - f_{\boldsymbol{\theta_2}}\|_{2,n} \lesssim \|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_2/\sqrt{J}. \tag{7.5}$$

Hence we can apply Theorem 2 for $J_n = C_1 n^{1/(2\alpha+1)}$, $M_n = M_0 > \|f_0\|_\infty$ as a constant, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}$ and $\epsilon_n = n^{-\alpha/(2\alpha+1)}\log n$. The conclusion then follows by Theorem 4 of Ghosal and van der Vaart [2007b]. The second part of the proof proceeds in the similar way as in Section 5.2. □

The above arguments apply for random covariates, too. Assume $Z_1, \ldots, Z_n$ has a marginal distribution g, then in the posterior calculation, we can absorb g in the dominating measure and denote it as G. Then Theorem 6 holds for random covariates while $\|\cdot\|_{2,n}$ is replaced by $\|\cdot\|_{2,n,G}$ and $\|\cdot\|_{2,n}^*$ is replaced by $\|\cdot\|_{2,n,G}^*$.

Remark 8. From (7.3), to get adaptation rate, we only need L_2-approximation for the true regression function. This implies that other basis expansions that has an L_2-approximation property can also be used for this model.

7.2 Binary regression

Bayesian methods have been used to study binary regression models for a while. The prior is commonly chosen as induced from a Gaussian process. A posterior consistency result was obtained by Ghosal and Roy [2006] while rates results were given by van der Vaart and van Zanten [2008]. Useful computational techniques were developed by Albert and Chib [1993] and Rasmussen and Williams [2006].

Assume that we have n independent observations $(Z_1, X_1), \ldots, (Z_n, X_n)$ from a binary regression model $\mathrm{P}(X = 1|Z = z) = 1 - \mathrm{P}(X = 0|Z = z) = f_0(z)$, where X's take values in $\{0, 1\}$ and Z's are either fixed or random covariates in some domain $\mathcal{Z}$. Given a link function $\Psi : \mathcal{Z} \to (0, 1)$, we can construct a prior on the regression function f_0 using spline functions as $f_{\boldsymbol{\theta}}(z) = \Psi\{\boldsymbol{\theta}^T \boldsymbol{B}(z)\}$. The likelihood function for (Z, X) can be written as

$$L_{\boldsymbol{\theta}}(z, x) = f_{\boldsymbol{\theta}}(z)^x (1 - f_{\boldsymbol{\theta}}(z))^{1-x} g(z), \tag{7.6}$$

where g is the marginal density of Z. When integrating out the posterior distribution, g will be canceled out. Therefore, in the following study, we can absorb g into the dominating measure, denoted it as G and remove g from the model (7.6). Define $p_{\boldsymbol{\theta}} = \Psi(\boldsymbol{\theta}^T\boldsymbol{B})^x(1-\Psi(\boldsymbol{\theta}^T\boldsymbol{B}))^{1-x}$ and $p_{f_0} = f_0^x(1-f_0)^{1-x}$, the following lemma states that if the link function Ψ satisfies certain conditions, then the the KL divergence can be controlled by the Euclidian distance (L_2-distance and supremum distance). Its proof is similar to Lemma 3.2 of van der Vaart and van Zanten [2008], and hence is omitted.

Lemma 2. *If Ψ possesses a bounded derivative ψ and the function $\psi/(\Psi(1-\Psi))$ is bounded, then for any $\boldsymbol{\theta_1}, \boldsymbol{\theta_2} \in \mathbb{R}^J$, $J \geq 1$ and $r > 1$, we have the following:*

$$\|p_{\boldsymbol{\theta_1}} - p_{\boldsymbol{\theta_2}}\|_r \lesssim \|\boldsymbol{\theta_1} - \boldsymbol{\theta_2}\|_{r,G} \tag{7.7}$$

$$\max\left\{K(p_{\boldsymbol{\theta_1}}, p_{f_0}), V(p_{\boldsymbol{\theta_1}}, p_{f_0})\right\} \lesssim \|\Psi(\boldsymbol{\theta_1}^T\boldsymbol{B}) - f_0\|_{2,G}^2. \tag{7.8}$$

In practice, some choices of link function Ψ can help us construct prior for $\boldsymbol{\theta}$ in an easy way. For example, if we choose Ψ as the identity link, then the prior of $\boldsymbol{\theta}$ can be defined on $(0,1)^J$. If we choose $\Psi(z) = z/(1+z)$, then the prior of $\boldsymbol{\theta}$ can be defined on all positive numbers.

Theorem 7. *Suppose the function $f_0 \in C^\alpha(0,1)$, $\alpha \leq q$ and the link function Ψ has a bounded derivative ψ while $\psi/(\Psi(1-\Psi))$ is bounded. Then the posterior distribution relative to the constructed prior converges at rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{1/2}$ with respect to $\|\cdot\|_{2,G}$ distance.*

Proof. Since $\psi/(\Psi(1-\Psi))$ is bounded, $p_{\boldsymbol{\theta}}$ is uniformly bounded below and above. Hence it is equivalent to work with $\|\cdot\|_{2,G}$ and Hellinger distance. Using Lemma 2 for $r = 2$ and Theorem 2 for $J_n \asymp n^{1/(2\alpha+1)}$, $M_n > \max\{\Psi^{-1}(f_0), f_0\}$, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}$ and $\epsilon = n^{-\alpha/(2\alpha+1)}(\log n)^{1/2}$, we can verify conditions in Theorem 4 of Ghosal and van der Vaart [2007b] in a similar way with the proof of density estimation. □

Remark 9. The above arguments imply that L_2-approximation is good enough to achieve adaptation. Hence we can also use other basis expansions that has an L_2-approximation property.

If f_0 is bounded away from 0 and 1, using Lemma 1, we can construct priors of $\boldsymbol{\theta}$ on $(0,1)^J$. This helps simplify the computation in a similar manner

with Section 5.3. Assume the identity link function and use Beta priors $\theta_i \overset{\text{ind}}{\sim} \text{Beta}(a_i, b_i)$ for some positive numbers a_i and b_i, the posterior mean of $f_0(z)$ is

$$\frac{\sum_{j=1}^{\infty}\Pi(j)\prod_{i=1}^{j}\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\sum_{i_0=1}^{j}\cdots\sum_{i_n=1}^{j}\int_0^1\prod_{k=1}^{j}\theta_k^{a_k-1}(1-\theta_k)^{b_k-1}\prod_{s=0}^{n}\theta_{i_s}^{X_s}(1-\theta_{i_s})^{1-X_s}B_{i_s}(Z_s)d\boldsymbol{\theta}}{\sum_{j=1}^{\infty}\Pi(j)\prod_{i=1}^{j}\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\sum_{i_1=1}^{j}\cdots\sum_{i_n=1}^{j}\int_0^1\prod_{k=1}^{j}\theta_k^{a_k-1}(1-\theta_k)^{b_k-1}\prod_{s=1}^{n}\theta_{i_s}^{X_s}(1-\theta_{i_s})^{1-X_s}B_{i_s}(Z_s)d\boldsymbol{\theta}},$$

where $Z_0 = z$. Define $I_{k,j,0} = \sum_{s=0}^{j}\mathbb{1}\{i_s = k\}$ and $I_{k,j,1} = \sum_{s=1}^{j}\mathbb{1}\{i_s = k\}$. The above equation can be simplified into

$$\frac{\sum_{j=1}^{\infty}\Pi(j)\prod_{i=1}^{j}\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\sum_{i_0=1}^{j}\cdots\sum_{i_n=1}^{j}\prod_{s=0}^{n}B_{i_s}(Z_s)\prod_{k=1}^{j}\frac{\Gamma(a_k+b_k+I_{k,j,0})}{\Gamma(a_k+\sum_{\substack{t=0,\ldots,j\\ i_t=k}}X_t)\Gamma(b_k+I_{k,j,0}-\sum_{\substack{t=0,\ldots,j\\ i_t=k}}X_t)}}{\sum_{j=1}^{\infty}\Pi(j)\prod_{i=1}^{j}\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\sum_{i_0=1}^{j}\cdots\sum_{i_n=1}^{j}\prod_{s=1}^{n}B_{i_s}(Z_s)\prod_{k=1}^{j}\frac{\Gamma(a_k+b_k+I_{k,j,1})}{\Gamma(a_k+\sum_{\substack{t=1,\ldots,j\\ i_t=k}}X_t)\Gamma(b_k+I_{k,j,1}-\sum_{\substack{t=1,\ldots,j\\ i_t=k}}X_t)}}.$$

7.3 Poisson regression

Consider a Poisson regression model $X_i \overset{\text{ind}}{\sim} \text{Poi}(f(Z_i))$, where f is a unknown monotonic function and Z_i's are covariates. For convenience, we assume Z_i's are fixed here, the random covariates case can be treated similarly as in Sections 7.1 and 7.2. Using a random series expansion, f can be modeled either as $f(z) = \boldsymbol{\theta}^T\boldsymbol{B}(z)$, where the prior of $\boldsymbol{\theta}$ is restricted on $(0,\infty)^J$ or through a link function $f(z) = g(\boldsymbol{\theta}^T\boldsymbol{B})(z)$, e.g., choose g as the exponential link function and allow $\boldsymbol{\theta}$ defined on $\mathbb{R}^J$.

The adaptation results can be obtained in a similar way with Section 7.1. Define an empirical measure $\mathbb{P}_n^Z = n^{-1}\sum_{i=1}^{n}\delta_{Z_i}$, the covariance matrix $\Sigma_{n,j}$, whose (k,l)-th element equals to $\int B_k B_l d\mathbb{P}_n^Z$ and $\|\cdot\|_{2,n}$ as the norm on $L_2(\mathbb{P}_n^Z)$. By applying the arguments in Section 7.1.1 of Ghosal and van der Vaart [2007b], we only need approximation results for L_2-distance. The posterior rates theorem is stated as follows

Theorem 8. *Suppose the true function $f_0 \in C^\alpha(\mathbb{R})$ for some $\alpha \le q$ and satisfies $L < f_0 < U$ for some constants $L, U > 0$. The priors are constructed as in*

Section 3.2. Then the posterior converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{1/2}$ *relative to* $\|\cdot\|_{2,n}$.

Remark 10. Since only L_2-approximation is needed, we can use other basis functions that has an L_2-approximation property as the expansion technique.

If we choose the identity link and let $\theta_i \overset{\text{ind}}{\sim} \text{Gamma}(a_i, b_i)$ for some positive numbers a_i and b_i, then the posterior mean of $f(z)$ is

$$\frac{\sum_{j=1}^{\infty}\Pi(j)\int_0^{\infty}\sum_{s=1}^{j}\theta_s B_s(z)\prod_{k=1}^{j}\frac{b_k^{a_k}\theta_k^{a_k-1}}{\Gamma(a_k)}\exp\{-\sum_{i=1}^{j}\theta_i(\sum_{k=1}^{n}B_i(Z_k)+b_i)\}\prod_{k=1}^{n}\{\sum_{i=1}^{j}\theta_i B_i(Z_k)\}^{X_k}d\boldsymbol{\theta}}{\sum_{j=1}^{\infty}\Pi(j)\int_0^{\infty}\prod_{k=1}^{j}\frac{b_k^{a_k}\theta_k^{a_k-1}}{\Gamma(a_k)}\exp\{-\sum_{i=1}^{j}\theta_i(\sum_{k=1}^{n}B_i(Z_k)+b_i)\}\prod_{k=1}^{n}\{\sum_{i=1}^{j}\theta_i B_i(Z_k)\}^{X_k}d\boldsymbol{\theta}}.$$

Denote $Z_0 = z$, $X_0 = 1$. Define $\boldsymbol{b} = (b_1, \ldots, b_J)^T$, $\boldsymbol{s}(k)$ as the k-th element of vector $\boldsymbol{s}$ and $\boldsymbol{B}(z)^{\boldsymbol{s}} = \prod_{i=1}^{j} B_i(z)^{s_i}$ for some index vector $\boldsymbol{s}$, the above quantity simplifies into

$$\frac{\sum_{j=1}^{\infty}\Pi(j)\sum_{\substack{\boldsymbol{s_0},\ldots,\boldsymbol{s_n}\in\mathbb{N}_0^j\\ |\boldsymbol{s_0}|=X_0,\ldots,|\boldsymbol{s_n}|=X_n\\ \boldsymbol{s}=\boldsymbol{s_0}+\cdots+\boldsymbol{s_n}}}\prod_{i=0}^{n}\frac{\{\boldsymbol{B}(Z_i)\}^{\boldsymbol{s_i}}}{\boldsymbol{s_i}!}\prod_{k=1}^{j}\frac{\Gamma(a_k+\boldsymbol{s}(k))b_k^{a_k}}{\Gamma(a_k)(b_k+\sum_{t=0}^{n}B_k(Z_t))^{a_k+\boldsymbol{s}(k)}}}{\sum_{j=1}^{\infty}\Pi(j)\sum_{\substack{\boldsymbol{s_1},\ldots,\boldsymbol{s_n}\in\mathbb{N}_0^j\\ |\boldsymbol{s_1}|=X_1,\ldots,|\boldsymbol{s_n}|=X_n\\ \boldsymbol{s}=\boldsymbol{s_1}+\cdots+\boldsymbol{s_n}}}\prod_{i=1}^{n}\frac{\{\boldsymbol{B}(Z_i)\}^{\boldsymbol{s_i}}}{\boldsymbol{s_i}!}\prod_{k=1}^{j}\frac{\Gamma(a_k+\boldsymbol{s}(k))b_k^{a_k}}{\Gamma(a_k)(b_k+\sum_{t=0}^{n}B_k(Z_t))^{a_k+\boldsymbol{s}(k)}}}. \quad (7.9)$$

7.4 Functional regression model

Spline functions are widely used to model functional data; see Cardot et al. [2003] for example. An asymptotic rates result of convergence was obtained in Hall and Horowitz [2007]. A Bayesian method based on splines is given by Goldsmith et al. [2011]. However, to the best of our knowledge, no results on posterior convergence rates for this model are yet available. We consider two types of functional regression model. The first one assumes only the covariates $Z(t)$ and the effects $\beta(t)$ depend on time t. The second one allows functional observations $X(t)$.

Suppose that the data that we observe are independent and identically distributed pairs $(Z_1, X_1), \ldots, (Z_n, X_n)$, where each Z is a square integrable random function defined on the unit interval and X's are scalers. Assume the residuals ϵ_i (i.i.d) follow a normal distribution with mean 0 and unknown variance σ^2. A functional regression model can be formulated as follows:

$$X_i = \int_0^1 Z_i(t)\beta(t)dt + \epsilon_i, \tag{7.10}$$

where $\beta(t)$ is the coefficient function we want to estimate.

The following assumptions are needed:

(E1) Smoothness: The function $\beta(t)$ is assumed to be α-Hölder.

(E2) Invertible condition: Assume $\mathrm{E}Z^2(t)$ is continuous and positive for every $t \in [0, 1]$.

Theorem 9. *Suppose that the true regression function β satisfies Conditions (E1) and (E2) and the prior is constructed as in Section 3.2 and* (7.1)*. Then, the posterior converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)} \log n$ relative to $\|\cdot\|_{2,Z}$, which is defined as $\|f\|_{2,Z}^2 = \int_0^1 f(t)^2 E(Z^2(t))dt$.*

Proof. Consider a B-spline expansion $\beta(t) = \sum_{k=1}^J \theta_k B_k(t)$. Denote $W_{ik} = \int_0^1 Z_i(t)B_k(t)dt$, then the model can be written as

$$X_i = \sum_{k=1}^J \theta_k W_{ik} + \epsilon_i. \tag{7.11}$$

Apply the same arguments as in Section 7.1, we can work on $\|\cdot\|_{2,Z}$ instead of Hellinger distance. Given Z, define P_β as the normal measure with mean $\int_0^1 Z_i(t)\beta(t)dt$ and variance σ^2. This allows us to bound the KL divergences using Cauchy-Schwarz inequality:

$$\begin{aligned} \max\big\{K(P_{\beta_0}, P_\beta), V(P_{\beta_0}, P_\beta)\big\} &\lesssim \frac{1}{\sigma^2}\mathrm{E}_Z\Big(\int_0^1 Z(t)\{\beta(t) - \beta_0(t)\}dt\Big)^2 \\ &\leq \frac{1}{\sigma^2}\|\beta - \beta_0\|_2^2\mathrm{E}\Big[\int_0^1 Z^2(t)dt\Big] \\ &\leq \frac{1}{\sigma^2}\|\beta - \beta_0\|_2^2. \end{aligned} \tag{7.12}$$

Hence we can apply Theorem 2 for $J_n = C_1 n^{1/(2\alpha+1)}$, $M_n = M_0 > \|\beta_0\|_\infty$ as a constant, $\bar{\epsilon}_n = n^{-\alpha/(2\alpha+1)}$ and $\epsilon_n = n^{-\alpha/(2\alpha+1)} \log n$. Therefore the conditions in Theorem 4 of Ghosal and van der Vaart [2007b] are verified. The proof is complete. □

Next, we consider a longitudinal type of functional model:

$$X_i(T_i) = Z_i(T_i)\beta(T_i) + \epsilon(T_i) \tag{7.13}$$

For each object i, we observe its response X_i at a random time T_i with some random covariate Z_i. We assume $Z_i \overset{\text{iid}}{\sim} Z$, $T_i \overset{\text{iid}}{\sim} T$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ for some finite probability distributions Z, T defined on the unit interval and some unknown variance σ^2. In order to obtain a similar argument with (7.12), we need an additional condition:

(E3) We assume T has a positive density function $g(t)$ at $[0, 1]$.

We have

$$\begin{aligned} \max\left\{K(P_{\beta_0}, P_\beta), V(P_{\beta_0}, P_\beta)\right\} &\lesssim \frac{1}{\sigma^2}\mathrm{E}_Z \int_0^1 Z^2(t)(\beta(t) - \beta_0(t))^2 g(t)dt \\ &\lesssim \frac{1}{\sigma^2}\mathrm{E}\{Z^2(t)\}\|\beta - \beta_0\|_2^2 \end{aligned} \tag{7.14}$$

By assuming the prior for σ satisfies (7.1) and using similar arguments in Theorem 9, we have the convergence theorem.

Theorem 10. *Suppose the true regression function $\beta(t)$ satisfy Conditions (E1)-(E3) and the prior is constructed as in Sections 3.2 and* (7.1)*. Then the posterior converges at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)} \log n$ relative to $\|\cdot\|_{2,Z}$.*

This rate coincides with the optimal rate obtained in Cai and Yuan [2011].

Remark 11. Note that the restriction of defining covariates on a compact interval can be dropped from Theorems 9 and 10 as we can project the support of covariates from $\mathbb{R}$ to the unit interval by a link function. Then the results proceed in a similar way with Section 5.2.

Remark 12. From (7.12) and (7.14), we observe that only L_2-approximation is needed. Hence other basis functions can be used as the expansion technique such as polynomial basis.

8 Extension to other function spaces

So far, our discussions are restricted to functions that belong to Hölder class. However, in applications, there is a lot of interest in studying a more general class of functions, e.g., Soblev and Besov spaces. We first give their definitions, then present the approximation results of B-splines in these function spaces.

Definition 1. *A Soblev space $L_p^\alpha[a,b]$ on a compact interval $[a,b]$ for $1 \le p \le \infty$ and a positive integer α is defined by:*

$$L_p^\alpha[a,b] = \{f : f^{(j)} \in L_p[a,b], j = 1, \dots, \alpha\} \tag{8.1}$$

with an associated norm $\|f\|_{L_p^\alpha[a,b]} = \sum_{j=0}^{\alpha} \|f^{(j)}\|_p$.

Definition 2. *A Besov space $B_{p,q}^\alpha[a,b]$ on $[a,b]$ for $1 \le p,q \le \infty$ and $\alpha > 0$ is the collection of all functions f such that*

$$|f|_{B_{p,q}^\alpha} = \Big[\int_0^\infty \{t^{-\alpha}\omega_r(f,t)_p\}^q \frac{dt}{t} \Big]^{1/q} < \infty, \tag{8.2}$$

where $r > \alpha$ is an integer and $\omega_r(f,t) = \sup_{|h|\le t} \|\Delta_h^r(f,\cdot)\|_p$ is called the modulus of smoothness of order r.

If $q = \infty$, then $|f|_{B_{p,q}^\alpha}$ is defined as $\sup_{t>0} t^{-\alpha}\omega_r(f,t)_p$ instead. Further define an associated norm on $B_{p,q}^\alpha[a,b]$ as

$$\|f\|_{B_{p,q}^\alpha} = \|f\|_p + |f|_{B_{p,q}^\alpha}. \tag{8.3}$$

The following two lemmas are taking from Chapter 6 of Schumaker [2007]. They describe approximation abilities of B-splines in these function spaces, which directly determines the posterior rates of corresponding Bayesian models.

Lemma 3. *Suppose $1 \le p \le q \le \infty$, for any $0 \le r \le \alpha - 1$ and $f \in L_p^\alpha[a,b]$, there exists a constant C and $\boldsymbol{\theta} \in \mathbb{R}^J$ such that*

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B}\|_{L_q^r[a,b]} \le C J^{r-\alpha-1/q+1/p} \|f\|_{L_p^\alpha[a,b]}.$$

In particular, if we choose $r = 0$ and $p = q$, the above inequality simplifies into

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B}\|_p \le C J^{-\alpha} \|f\|_p.$$

Lemma 4. *For any function $f \in B^{\alpha}_{p,q}[a,b]$, $1 \le p \le p' \le \infty$ and $1 \le q, q' \le \infty$, $0 < \tau < \lfloor \alpha \rfloor$, there exists a constant C and $\boldsymbol{\theta} \in \mathbb{R}^J$ such that*

$$\|f - \boldsymbol{\theta}^T \boldsymbol{B}\|_{B^{\tau}_{p',q'}[a,b]} \le C J^{\tau - \alpha - 1/p' + 1/p} \|f\|_{B^{\alpha}_{p,q}[a,b]}.$$

If we can bound the KL divergence by an appropriate norm in the new function spaces, then the posterior rates calculation will be very similar with what we have done in the previous discussion. For example, take $p = 2$ in Lemma 3, we can show that the posterior rate results in Theorem 4 remains the same for $f_0 \in L^{\alpha}_2[a,b]$ and the use of log link function. This is true because

$$\begin{aligned} h^2(f_0, f_{\boldsymbol{\theta}}) \le V(f_0, f_{\boldsymbol{\theta}}) &\lesssim \|\log f_0 - \boldsymbol{\theta}^T \boldsymbol{B}\|_2^2, \\ K(f_0, f_{\boldsymbol{\theta}}) \le 2h^2(f_0, f_{\boldsymbol{\theta}}) \Big\| \frac{f_0}{f_{\boldsymbol{\theta}}} \Big\|_{\infty} &\lesssim \|\log f_0 - \boldsymbol{\theta}^T \boldsymbol{B}\|_2^2, \end{aligned} \tag{8.4}$$

where $f_{\boldsymbol{\theta}} = \exp\{\boldsymbol{\theta}^T \boldsymbol{B}\} / \int \exp\{\boldsymbol{\theta}^T \boldsymbol{B}\}$ is the approximating density function, which is lower bounded by a multiple of $\exp\{-c\|\boldsymbol{\theta}\|_{\infty}\}$ for some positive constant c. Similar results apply for regression, spectral density models, too.

For Soblev spaces, the restriction is that Lemma 3 only considers the case when the smoothness parameter α is an integer. It is not clear whether such approximation result remains valid for non-integer values of α. For Besov spaces, notice that there is an extra power J^{τ} in Lemma 4, where τ is strictly greater than 0. This suggests only a sub-optimal rate $\epsilon_n = n^{-\alpha/(2\alpha+1)+\tau}(\log n)^c$ $(\tau, c > 0)$ can be achieved by our methods, though τ can be chosen arbitrarily close to 0. This is because we only need to bound the approximation error under Euclidean distances, extra approximation results on derivatives are not necessary.

9 Numerical results

We illustrate the use of conjugate prior structure as described in Section 5.3. Following Lenk [1991], we generate 50 samples from a mixture density of exponential and a normal distribution

$$f_0(x) \propto \frac{3}{4} 3e^{-3x} + \frac{1}{4}\sqrt{32/\pi} e^{-32(x-0.75)^2}. \tag{9.1}$$

We implement the random series prior using quadratic B-splines ($q = 3$) and choose a geometric prior Geo(.15) for J restricted between 5 and 12. The

lower truncation ensures a minimum number of terms in the series expansion while an upper truncation is necessary to carry out the actual computation using a computer. For $\boldsymbol{\theta}$, we use a Dirichlet distribution with parameters $a_1, \ldots, a_J = 1$. Instead of iterating all 3^{50} possible permutation of indices to get equation 5.6, we randomly sample $N = 1000$ of them and take the associated average values. We obtained density estimates on 1000 grid points in the unit interval. The maximum Monte-Carlo standard error of the estimates is 0.12 calculated by Delta method. The computation takes about 15 minutes on a 2.20 GHz machine. In contrast, it takes 1 hour to run an RJMCMC method on the same problem. We compare our results with the use of Gaussian process prior in Tokdar [2007]. Our method has a mean squared error (MSE) 0.076 to the true density while the MSE for Gaussian process prior is 0.111. Figure 1 shows that random series prior has a comparable performance with Gaussian process prior.

References

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, 88:669–679, 1993.

A. Babenko and E. Belitser. Oracle convergence rate of posterior under projection prior and Bayesian model selection. *Math. Methods Statist.*, 19:219–245, 2010.

S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *J. Roy. Statist. Soc. Ser. B*, 70: 825–848, 2008.

E. Belitser and S. Ghosal. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31:536–559, 2003.

L. Birgé. Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincar Probab. Statist*, 42:273325, 2006.

T Cai and M. Yuan. Optimal estimation of the mean function based on discretely sampled funcitonal data: phase transition. *Ann. Statist.*, 39:2330–2355, 2011.

H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Stat. Sinica.*, 13:571–591, 2003.

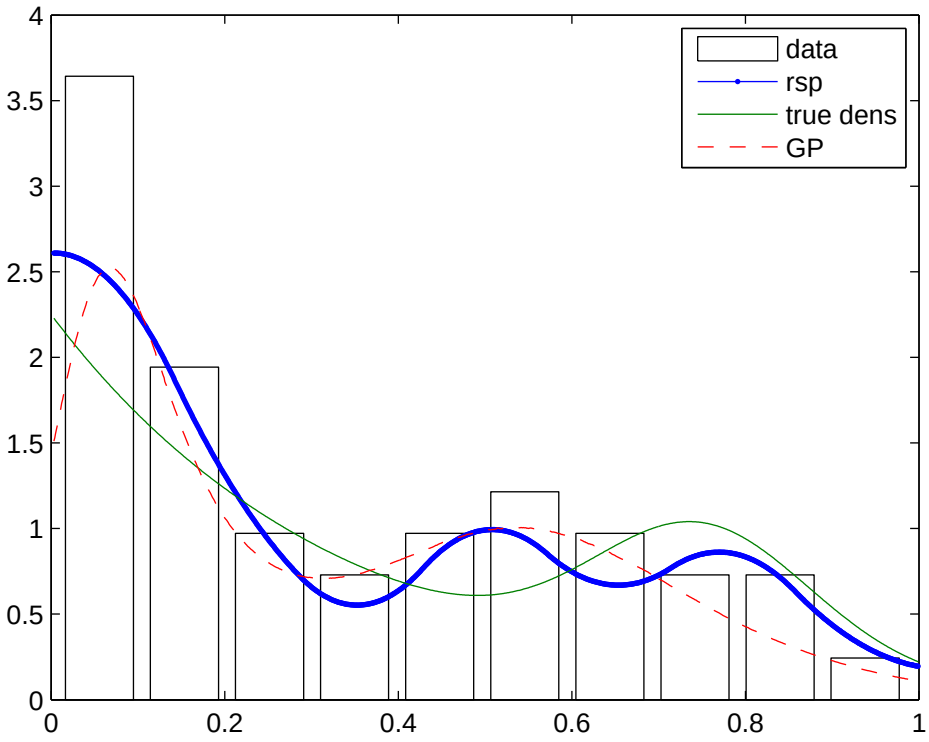

Figure 1: Approximated Bayesian estimates using random series prior (rsp) and Gaussian process prior (GP)

I. Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.*, 2:1281–1299, 2008.

I. Castillo. A semi-parametric Bernstein-von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields*, 152:53–99, 2012.

T. Choi and M. J. Schervish. On posterior consistency in nonparametric regression problems. *J. Multivar. Anal.*, 98:1969–1987, 2007.

N. Choudhuri, S. Ghosal, and A. Roy. Bayesian estimation of the spectral density of a time series. *J. Amer. Statist. Assoc.*, 99:1050–1059, 2004a.

N. Choudhuri, S. Ghosal, and A. Roy. Contiguity of the Whittle measure for a Gaussian time series. *Biometrika*, 91:211–218, 2004b.

C. de Boor and J. W. Daniel. Splines with nonnegative b-spline coefficients. *Math. Comp.*, 28:565–568, 1974.

Carl de Boor. *A Practical Guide to Splines*. Springer, 2001.

R. de Jonge and J. H. van Zanten. Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. Preprint, 2012.

M. Gasparini. Bayesian density estimation via dirichlet density processes. *J. Nonparametr. Stat.*, 6:355–366, 1996.

S. Ghosal. Convergence rates for density estimation with Bernstein polynomials. *Ann. Statist.*, 29(5):1264–1280, 2001.

S. Ghosal and A. Roy. Posterior consistency of Gaussian process prior for nonparametric binary regression. *Ann. Statist.*, 34:24132429, 2006.

S. Ghosal and A. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(3):697–723, 2007a.

S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35:192–223, 2007b.

S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27(1):143–158, 1999.

S. Ghosal, J. K. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

S. Ghosal, J. Lember, and A. van der Vaart. On Bayesian adaptation. *In Proceedings of the Eighth Vilnius Conference on Probability Theory and Mathematical Statistics, Part II (2002)*, 79:165–175, 2003.

S. Ghosal, J. Lember, and A. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89, 2008.

E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Ann. Probab.*, 37:1605–1646, 2009.

J Goldsmith, Matt P. Wand, and Ciprian Crainiceanu. Functional regression via variational Bayes. *Electron. J. Stat.*, 5:572602, 2011.

Peter Hall and Joel L. Horowitz. Methodology and convergence rates for functional linear regression. *Ann. Statist.*, 35:70–91, 2007.

J. S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral Methods for Time-Dependent Problems*. Cambridge University Press, 2007.

T.-M. Huang. Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32:1556–1593, 2004.

Dunham Jackson. *The Theory of Approximation*. AMS Colloquium Publication Volume XI, New York, 1930.

P. J. Lenk. The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.*, 83:509–516, 1988.

P. J. Lenk. Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78:531–543, 1991.

T. Leonard. Density estimaiton, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B*, 40:113–146, 1978.

H. Lian. On posterior distribution of Bayesian wavelet thresholding. *J. Statist. Plann. Inference*, 141:318–324, 2011.

G. G. Lorenz. *Bernstein Polynomials*. Univ. Toronto Press, 1953.

S. Petrone. Random Bernstein polynomials. *Scand. J. Statist.*, 26:373–393, 1999a.

S. Petrone. Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist*, 27:105–126, 1999b.

M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian white noise. *Problems Inform. Transmission*, 16:120–133, 1980.

C. E. Rasmussen and C. K. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

V. Rivoirard and J. Rousseau. Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.*, 2012. To appear.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Statist. Soc. Ser. B*, 71:319–392, 2009.

L. Schumaker. *Spline Functions: Basic Theory*. Cambridge University Press, 2007.

C. Scricciolo. Convergence rates for Bayesian density estimation on infinite-dimensional exponential families. *Ann. Statist.*, 34:2897–2920, 2006.

X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29:687–714, 2001.

C. J. Stone. Optimal global rates of convergence for nonparametric regressioin. *Ann. Statist.*, 10:1040–1053, 1982.

C. J. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18: 717–741, 1990.

S. T. Tokdar. Towards a faster implementation of density estimation with logistic gaussian process priors. *J. Comput. Graph. Statist.*, 16:633–655, 2007.

S. T. Tokdar and J. K. Ghosh. Posterior consistency of logistic gaussian process priors in density estimation. *J. Statist. Plann. Inference*, 137:34–42, 2007.

Y. Truong, C. Kooperberg, and C. Stone. *Statistical Modeling with Spline Functions: Methodology and Theory*. Springer, 2005.

A. van der Vaart and H. van Zanten. Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.*, 1:433–448, 2007.

A. van der Vaart and H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36:1435–1463, 2008.

A. van der Vaart and H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37 (5B):2655–2675, 2009.

P. Whittle. Curve and periodogram smoothing (with discussion). *J. Roy. Statist. Soc. Ser. B*, 19:38–63, 1957.